Research on the relationship between residents' fitness and urban space based on social media data: a case study of Changsha Zhuzhou Xiangtan area

Yu CHEN[1,2], Shaoyao He[1], Jianmin Zhao[1], Yuanqiang Tang[1]

[1]Hunan University, Changsha, Hunan, China

[2]Hunan City University, Yiyang, Hunan, China

Keywords: social media data, LDA, social behavior, nuclear density analysis

Abstract: Under the background of big data, social media data, as an important spatial data, plays a very important role in the research of urban residents' spatiotemporal behavior. In this paper, through crawling the data of sina Weibo with text and location information in Changsha Zhuzhou Xiangtan area, this paper selects more than 10000 pieces of data related to urban residents' fitness behavior by manual tagging and machine learning. After data cleaning, LDA method is used to analyze the text and extract the subordinate topics related to fitness behavior Nuclear density analysis and fishing net analysis reveal the spatial characteristics of healthy sports and the social, economic, cultural and other factors that affect their distribution. It is of great significance for government departments and planning departments to examine and analyze the spatial structure of healthy cities in rapid development and guide the spatial planning and construction of healthy sports and industrial selection.

1. Introduction

With the highly developed economy, the increasingly mature society and the influence of humanistic trend of thought, urban development is no longer only pursuing economic goals, and the urban development concept with people as the center and improving the quality of life as the goal has become the consensus of studying urban social life. Improving national fitness consciousness, strengthening physical health and improving physical quality has been paid more and more attention by the public, and the public is participating in the fitness exercise more and more [1]. Based on the text content and picture information, we can analyze the topics that users are interested in. The popularity of users on this platform can be determined based on the number of fans and the number of times they are forwarded. Based on time stamps, we can analyze the daily active time period and activity degree of users [2]. By collecting this kind of demographic data and analyzing it in combination with factors such as time, society and geographical space, we can intuitively observe human behavior patterns. For example, people from different parts of the world have different daily sleep patterns and different ways of spending winter and summer vacations.

In this paper, Changsha-Zhuzhou-Xiangtan area is selected as the research object, and the data set of Sina Weibo is used for empirical research. By collecting VGI data of public fitness, the current situation of residents' fitness path in urban areas is studied, and the relationship between public fitness and the distribution of urban park green space is explored in combination with the spatial layout of urban park green space. This paper reveals the inherent spatial structure of the city from the perspective of human activity dynamics, and provides decision support for guiding the spatial planning and construction of healthy sports and the choice of industries.

2. Preprocessing of Interest Point Data and Social Media Sign-in Data

In this paper, the official sign-in data of Sina Weibo from 2018 to 2019 are used, including 1893041 sign-in records and 13785 points of interest (POI). Like any other LBSN, users connect with applications by signing in and interact with other people in the network.

Figure 1 shows the track data in spatio-temporal database and sign-in data in social network formally. The time interval between adjacent points in the same track is usually short [3], while the

  DOI: 10.25236/iwass.2021.039

time interval between adjacent sign-in records in social networks is usually large, and some even last for several months [4].

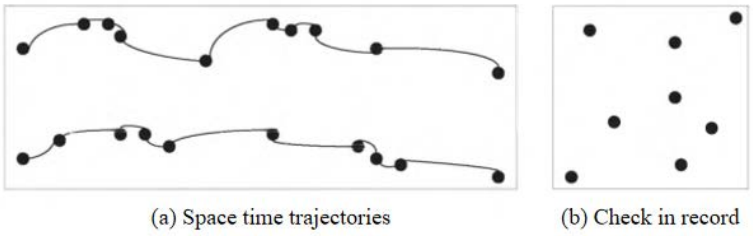

Figure 1 Track data and sign-in data

With the popularity of computers in many fields, people are active on many different platforms at the same time, and the development of the whole industry presents a fragmented trend. Connecting the same users from different platforms can obtain more abundant data, thus analyzing user behavior in more detail. Time stamp-based clustering can find out which time points users are active every day[5]. These check-in records are usually highly discrete, because on the one hand, users will not continuously share the status in a very short time, and on the other hand, some users will not attach check-in data to each shared status for privacy and security reasons.

3. Research Model

3.1. Social media data analysis framework

With the development of society and the continuous advancement of the national fitness program, the fitness needs of urban residents in China are constantly developing and changing, which also has a certain impact on the dissemination of sports information in social media. According to the 2019 Survey Bulletin on the Status Quo of Physical Exercise Participation of Urban and Rural Residents in China, the top three purposes of physical exercise participation of urban and rural residents in China are weight loss, recreation and disease prevention and treatment in turn (see Figure 2).

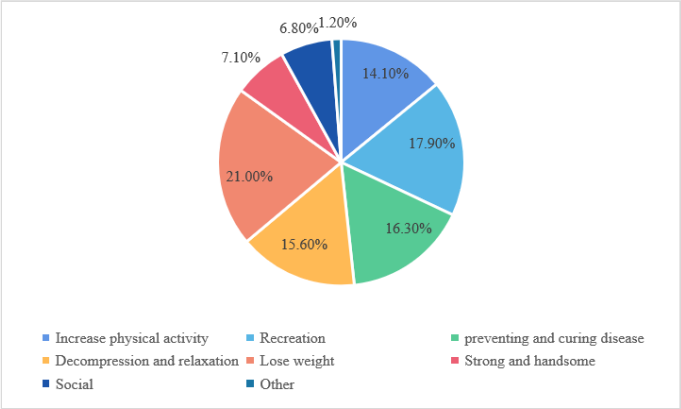

Figure 2 Proportion of urban and rural residents participating in physical exercise in 2019

It can be seen from the table that the fitness needs of "recreation", "weight loss" and "prevention and treatment" are increasing. While social media is constantly affecting the lives of modern urban residents, more fitness people are shifting their interest points to social media with the same recreation and decompression functions.

All kinds of information related to sports and fitness on social media come from every registered user. According to the data center of Sina Weibo, on November 1, 2019, there were 306 million users of general sports on Sina Weibo, of which 16 million were actually interested in sports [6-7]. The gender distribution ratio of actual sports interest users is 67.96% for men and 32.04% for

women. The educational background distribution ratio is 76% for undergraduate degree or above, 18% for senior high school, 5% for junior high school and 1% for primary school. See Table 1 for gender distribution of age groups.

Table 1 Gender distribution of users with actual sports interests at different ages

Age	Male ratio	Proportion of women	Age	Male ratio	Proportion of women
Under 18 years old	7.3%	4.6	30-34 years old	7.4	2.4
18-24 years old	27.1	12.7	35-40 years old	3.1	1.2
25-29 years old	21.3	8.6	Over 40 years old	2.8	1.5

City is an objective physical form. City users are all kinds of people. According to the role in the city, it can be divided into tourists, white-collar workers or primary school students [8]. For example, clustering based on spatial points can find the areas where users are frequently active, and clustering based on time stamps can find out which time points users are active every day. Generally, the factors involved are economic development level, industry proportion, transportation developed level, population size and population education level (or their representative information technology level), etc. The functions used are multivariate linear form and power function product form. Figure 3 describes the general framework of spatial analysis. The first stage of data collection is to download data from Sina Weibo. The next stage includes two parts: cleaning and analysis of LBSN data. Document topic generation model LDA method is used for text analysis, and the subordinate topics related to fitness behavior are extracted. Statistical analysis (sign-in probability) and KDE data visualization are used in the analysis stage, and ArcGIS (www.arcgis.com) is used to generate density map.

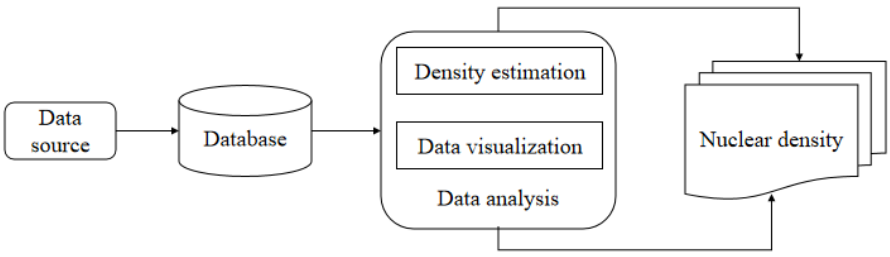

Figure 3 Data analysis framework

The interaction between Sina Weibo and urban residents' fitness needs is mainly reflected in two aspects: running and weight loss and fitness. With the heated discussion of sports and fitness topics on Sina Weibo, users' interest points gradually shift from weight loss topics to sports and fitness topics with a higher degree of actual participation. Some users will not attach check-in data in each sharing state for privacy and safety reasons. Therefore, the city center has higher requirements on the level of identifiability than the district center and community center. In other words, by improving the spatial identifiability at all levels of the city, the effect of active crowd gathering can be achieved, and social interaction and vibrant city life can be promoted.

3.2. Analytical method

KDE is a nonparametric method to estimate density by randomly selecting samples from data [9]. KDE calculates smooth distribution by eliminating local noise to some extent, and this method minimizes errors by providing nonparametric probability distribution with optimal bandwidth. Considering that different POIs should be closely connected with each other, the number of consecutive visits is used as the edge right. However, when people perceive the urban space, they often go from the roof of the building, the color of the street trees, to the residential area, a favorite

restaurant, and finally after a period of time, they can understand a certain area. Different from the discrete sign-in data in social networks, the spatio-temporal trajectory data is usually continuous, because the location points that make up the trajectory are automatically and continuously sampled by GPS devices, and are not transferred by the subjective will of moving objects.

E represents a series of historical data, in which $e^j = \langle x, y \rangle$ represents the position coordinates of a place, and j satisfies $1 \le j \le n$ for an individual. h_j represents the euclidean distance from the k th nearest neighbor e^j in the training data. The kernel density function $K(\cdot)$ is used to estimate the binary density function of the data, and the following bivariate KDE formula is derived [10]:

$$f_{KD}(e \mid E) = \frac{1}{n} \sum_{j=1}^{n} K_{h^j}(e, e^j)$$

In which: f_{KD} represents the obtained density estimate.

Residents' fitness exercise has the direct purpose of enhancing physical and mental health. Understanding the distribution characteristics of residents' sports paths in urban environment can better strengthen the interaction between people's fitness exercise and park green space, and improve residents' enthusiasm for participating in fitness exercise. However, in suburban districts far from the city center, a certain social gathering is formed in the center of the group. Considering the difference between track data and sign-in data, it is difficult to calculate the similarity between users directly based on these two kinds of data. In social science research, Pearson correlation coefficient is usually calculated first to observe whether the correlation between variables is significant, and then it is decided whether to use regression analysis to explore the predictive power and explanatory power of factors according to the value of correlation coefficient. Spatial grouping under a certain function makes it more difficult for people to accurately identify different levels of information in cities.

4. Result Analysis

The society of information explosion brought by the influence of Internet life has also changed the traditional way of understanding cities. The ways and means of identifying cities show new features of first-in-first-out, and also put forward different demands for the identifiability of space. First, it is necessary to divide the annual check-in data into time periods and construct the network respectively; Second, it is necessary to set a time threshold. For the continuous sign-in relationship in the user's track that is larger than the threshold, no edge is constructed or weighted, and only the continuous sign-in relationship within the threshold is considered. Therefore, urban design, which is influenced by the values of the city's identifiability, is a process of improving the quality of the city's open space and clarifying its spatial characteristics. Sign-in records in social networks are highly discrete and published subjectively by users, and each record contains important semantic information.

In order to improve the accuracy of text analysis, this paper firstly cleans the text data in the sign-in records to remove some noise data, then uses the classic TF-IDF algorithm in text mining to find out 20 places frequently visited by each type of people [11], and extracts the user sign-in trajectory network containing these 20 places. Finally, it uses the modular algorithm in the community discovery algorithm to analyze the trajectory network and get the clustering results of interest points (Figure 4).

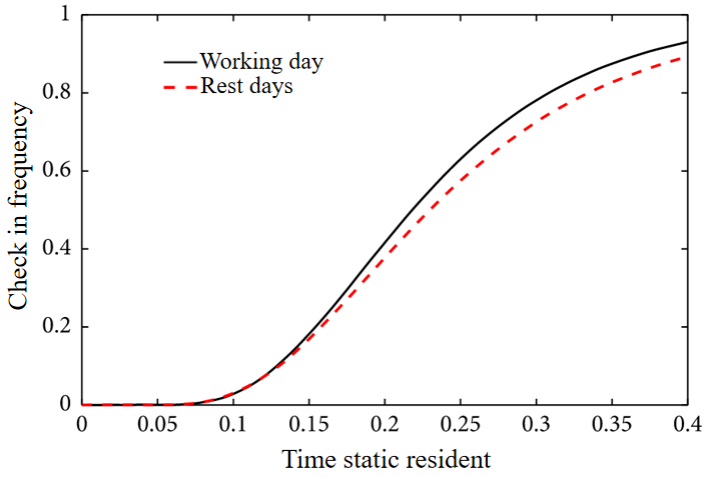

Figure 4(a)

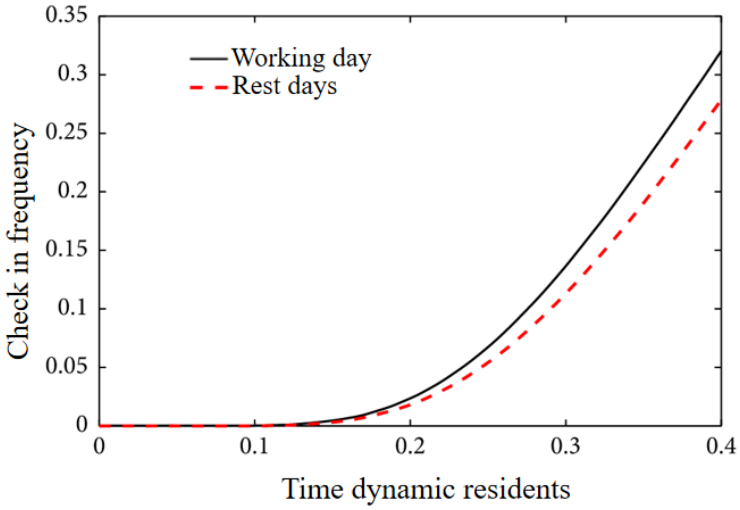

Figure 4(b)

Figure 5 Sign-in changes of different types of people in different time periods

In addition to check-in time and coordinates, the check-in record also contains text information. By studying and analyzing the text data and trajectory data of the user's check-in, we can find the interest points of the user's check-in and the potential relationship between the interest points. According to fig. 4(a), because static residents are permanent residents in Shenzhen, their activity trajectory network is complex and cross-regional.

There are differences in residents' fitness paths. On the one hand, there are differences in personal attributes such as gender, age, health condition, etc., on the other hand, there are differences in residents' family location and accessibility to parks, which all result in a wide distribution of residents' fitness paths. Taking income as an example, low-income people's choice of residence is often limited by housing price, commuting time and other factors. Therefore, the main city center is often a gathering place for low-income and middle-income people, which is related to the reasons that the main city can provide a large number of jobs, and the middle-income people are mainly young white-collar workers, and have high demand for commuting convenience. Because many social platforms exist at the same time, and the emphasis of each platform is different, users often don't concentrate on a specific platform. This shows that the node attraction connotation of social media check-in system is obviously different from other spatial interactive systems, and tourism competitiveness has replaced population size as the main influencing factor [8].

There are many factors that affect residents' fitness space. Urban land use, government planning and policies can change the development pattern, spatial structure and socio-economic layout of cities from a macro perspective; The degree of economic development and living habits of cities will change the residents' activity space from different angles and in different ways (Figure 5).

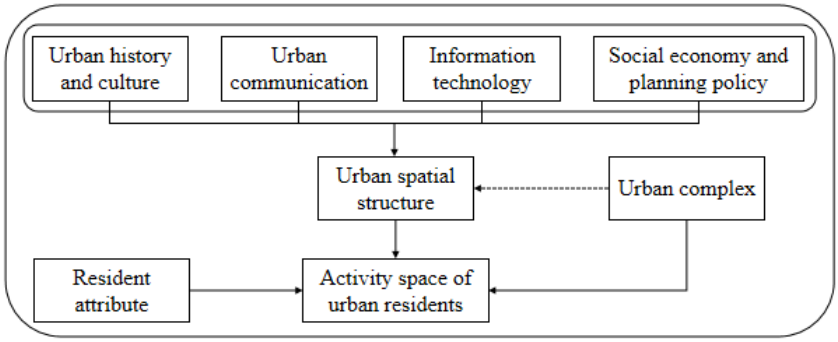

Figure 5 Analysis of influencing factors of residents' fitness space

Community activity distribution can be regarded as community size distribution weighted by check-in times. Analysis of α value shows that the weighted result can enhance the heterogeneity of the original distribution. Professional sports and fitness practitioners and related institutions and enterprises are willing to gain higher credibility through certification to attract fans' attention and achieve more communication value; The large-scale spatial shift of users makes the area of a given data set too large, which leads to the decline of the check-in data density. That is to say, the degree of "interactive enthusiasm" on social media should be able to positively predict the positive changes of individual's "fitness frequency", "fitness mode" and "fitness diet" in real life. The correlation analysis of these four variables gives the following correlation coefficient table (see Table 2).

Table 2 Pearson product moment correlation coefficient

	Meet other people's fitness show	Increase in fitness frequency	Fitness is more suitable	Fitness diet is more reasonable
Meet other people's fitness show	1			
Increase in fitness frequency	.227**	1		
Fitness is more suitable	.301*	.721**	1	
Fitness diet is more reasonable	.207*	.625**	.708*	1

* *. was significantly correlated at .01 level (bilateral).

Individuals who contact others through social media with high frequency of fitness show have higher frequency of fitness in real life than those with low frequency, and the fitness mode they choose is more suitable for themselves, and the diet plan for assisting fitness is more reasonable. This is because given a historical data set with large volume and high density, the real behavior patterns of users can be found more accurately based on these data, which is beneficial to find the actual connected user pairs. In addition, the traffic in the main city is convenient, and walking, cycling and public transportation are the main ways for residents to travel. The cycle of exercise can show the use frequency of park green space to a certain extent. For example, when the park is used frequently, it shows that the park is popular with the public and the park environment is attractive to the fitness crowd. Only the first-class urban centers have significant clustering characteristics in the service efficiency of social interaction among the three urban spaces. However, the secondary and tertiary open space centers are lack of attracting enough social activities.

5. Conclusion

The interactive content between social media and urban residents' fitness needs has changed from competitive sports to mass sports and mass sports. In essence, this process is a process of measuring the input cost and expected benefit of behavior. "Perceived personal obstacles" is the cost of personal perceived fitness behavior, and the lower the obstacles, the lower the cost; The "alternative experience" will affect the individual's expectation of their own fitness results, that is, the expected income. There are still many open space areas in cities that need to be further improved. At the planning and design level, it is clear that various types of open spaces need to strengthen the activity intensity and contact between citizens and open spaces. Therefore, it is of great significance to link the overall layout of park green space with residents' daily fitness activities, constantly improve the development layout of urban park green space and form an urban green space system that can promote residents' fitness activities.

Acknowledgements

This study was financially supported by the National Natural Science Foundation of China (grant NO.:51978250).

References

[1] Zhou Yan, Li Yanxi, Huang Yueying, et al. Urban population classification and activity characteristics analysis based on social media data [J]. Journal of Geoscience, 2017, 19(009):1238-1244.

[2] Zhang Depeng, Pan Guanglei, Zhang Hui, et al. Analysis of spatial and temporal dynamic changes of Hangzhou city based on check-in data [J]. Science and Technology Economic Guide, 2019, 27(14):26-28.

[3] Yu Xuesong, Jia Tao. Scale-free and hotspot analysis of spatial network and its community based on social media check-in data [J]. Chinese Science and Technology Papers, 2018, v.13(15):116-123.

[4] Zhang Qinglan, Cheng Gang, Zhang Yifan, et al. Study on the characteristics of urban contact network based on social media data-taking Yangtze River Delta urban agglomeration as an example [J]. Resources Guide, 2019, 000(002):29-31.

[5] Saqib, Ali, haidery, et al. geospatial analysis of Sina Weibo data based on kernel density estimation: a case study of Shanghai [J]. electronic measurement technology, 2019, v.42; No.329(21):37-43.

[6] Feng Rong, Liu Lu, Ma Di Xiang, et al. A dimension of street space quality of odor landscape [J]. Times Architecture, 2017, 000(006):18-25.

[7] Huai Songyao, Chen Zheng, Liu Song. Study on the quality of urban public space based on new data and new technology [J]. Urban Architecture, 2018(6): 12-20.

[8] Chen Zihao, Zhang Yi, Liu Yu, et al. Perception of urban tourism activities based on social media-taking Suzhou as an example [J]. Geography and Geographic Information Science, 2020(2):54-61.

[9] Xie yongjun, Peng Xia, Huang Zhou, et al. image perception of hot spots in Beijing based on microblog data [J]. advances in geographical science, 2017, 36(009):1099-1110.

[10] Huai songyao, Chen Zheng, Liu song. the quality of urban public space based on new data and new technologies [j]. urban architecture, 2018, 000(006):12-20.

[11] Li Canghai, Xu yitie, Luo chunhai, et al. spatial analysis of microblog information diffusion [J]. complex system and complexity science, 2017(3):75-84.